Data Quality and Missing Data in Patient-Centered Outcomes Research Using EMR/Claims Data Meeting Summary

Overview

On December 10, 2015, PCORI convened a multi-stakeholder workgroup to discuss current research and explore opportunities to help resolve key issues and challenges presented by missing data and data quality using Electronic Medical Record (EMR)[1] and Claims Data. The meeting focused on commonly encountered problems, existing solutions, and refined the research agenda needed to improve data quality.

Workgroup members included biostatisticians, informatics experts, epidemiologists, representatives from PCORnet networks, PCORI staff, and researchers from industry, universities, healthcare organizations, and federal agencies. The meeting was held at PCORI's offices in Washington, DC.

Related Information

- PCORI Methodology Commitee
- PCORnet, the National Patient-Centered Clinical Research Network

The Patient-Centered Outcomes Research Institute (PCORI) is an independent organization created to help people make informed healthcare decisions.

1828 L St., NW, Suite 900
Washington, DC 20036
Phone: (202) 827-7700
Fax: (202) 355-9558
Email: info@pcori.org

Introduction

PCORI's Acting Chief Science Officer, Dr. Harold Sox, opened the meeting with an overview of PCORI and PCORnet and their mission to help inform healthcare decisions. He described PCORI's interest in missing data as it relates to research in the context of clinical practice, as well as PCORI's interests in developing research questions and best practices for reducing and handling missing data.

Cynthia Girman, Member of PCORI's Methodology Committee (Ex-Officio member of the Advisory Panel on Clinical Trials), then provided additional background on the meeting. She described the evolution of EMRs from paper charts over the past few decades. EMR databases contain diagnoses, physician notes, vital signs, lab work, pharmacy (Rx filled), and claims linked to EMRs. Missing data may occur at any

[1] "**Electronic medical records (EMRs)** are a digital version of the paper charts in the clinician's office. An EMR contains the medical and treatment history of the patients in one practice. **Electronic health records (EHRs)** are designed to reach out beyond the health organization that originally collects and compiles the information. They are built to share information with other health care providers, such as laboratories and specialists, so they contain information from all the clinicians involved in the patient's care."
http://www.healthit.gov/buzz-blog/electronic-health-and-medical-records/emr-vs-ehr-difference/

point from initial clinical observation and recording to data extraction into a "clean" analytic data set, and sometimes a researcher may not even know that data are missing. Patient-level EMR data linked with patient-level claims are a unique and rich data source, so understanding the issues of missing data in clinical research networks is pivotal for generating and interpreting findings derived from these networks. Girman reiterated that the focus of the meeting would be to identify gaps and methodological issues with missing and inaccurate data, identify where such gaps typically occur in EMR/claims for PCOR studies, identify what research might be pursued to address such gaps in methodology, and to determine the potential for meaningful research collaborations.

Cross-Sectional Missing Data

Workgroup member Michael Kahn from UC Denver began the presentations. He discussed health data quality and variation in the terms used to describe it. He provided several examples of terminology used to help understand data quality—including believability, consistency, validity, and accuracy. Kahn raised the need for harmonizing terms in order to harmonize methods to handle missing data—which is ultimately needed to increase the transparency and trust in research findings. As part of his work under a PCORI Methods Contract (#5581), a community of clinical data quality experts examined the diversity in use of data quality terms, and ultimately determined that the existing heterogeneity of data quality descriptions could be harmonized under three major categories of data quality: completeness, fidelity, and plausibility. Kahn observed that there is not an explicit category of missingness, that rather it is considered a special case of completeness. He asked if the group was more concerned about documenting the scope of missing data, or did we also need to know the underlying cause of missingness?

The discussion that followed yielded the following points and comments:

- The cause of missing data is imperative to understand, as that information is needed to understand how to improve completeness and the potential biases embedded in missing data. It is also important to understand the incentives of those entering data.
- A component of missing data is that it must be relevant to the intention for analysis (called "fitness for use"). An example of this would be data missing on pediatric populations in a study focused exclusively on senior care. Collaborators at the health system level need to be aware of system procedures and incentives.
- The distinction between missing, true zero, and null often brings up many analytic issues. Similarly, understanding the presence and cause of missing data on the denominator population is important for estimating the impact that missing data has on analysis.
- A potential research area that might be further explored concerns sensitivity analysis and simulation to better understand the effects of differing degrees of missing data. The field could devise tests that inform researchers if the degree of missingness is such that results will be influenced and, hence, whether it is worth addressing missing data in the research question under consideration.
- "Edit" is a part of analysis (even though some would not consider this so). There is a lot of software dealing with missing data, but there is very little software for edit and imputation.
    - There was a lack of consensus on what "edit" defined. For example, is "edit" the process of putting back in data that are missing using some other data source that provides the

missing value? Or is "edit" somebody making an educated guess about what the missing data element must be based on other values present in the data set?

- Given that there are existing approaches to missing data in the claims setting (see description of next workshop session), can we utilize a similar paradigm/approach in the EMR setting?
- For data networks, should missing values be imputed at the global level or should they be done locally at each node of the network? Global imputation means you impute once across the entire network, and utilize that dataset for all subsequent analyses versus local imputation means impute separately for each analysis, or indeed the individual researcher imputes versus the researcher being provided one (public use?) imputed dataset.
    - Several at the meeting were against the idea of global imputation as it would be unrealistic to think we could impute all missing data at all sites.
- Should imputation be done overall, or for each specific research question? The challenge of PCORnet is that it is a general utility that is also intended to be used in very specific ways. The PCORnet Common Data Model's endorsement for "leaving the data as they are in the native system, warts and all" could imply that PCORnet would not want imputed values stored next to "real" values in the CDM.
- Information on data "cleaning" must be made more explicit, and guidance is needed on how to describe data cleaning?

Related resources:

- Kahn MG, Brown J, Chun A, Davidson B, Meeker D, Ryan P, et al. Transparent Reporting of Data Quality in Distributed Data Networks. eGEMs (Generating Evidence & Methods to improve patient outcomes) [Internet]. 2015 Mar 23;3(1). Available from: http://repository.academyhealth.org/egems/vol3/iss1/7

Appropriate Use of Claims Data

Alan Brookhart, from UNC-Chapel Hill, then presented on Missingness and Error in Variables Based on Pharmacy Claims & Hospitalization/Diagnosis Data. He described the characteristics of pharmacy claims data, which capture medications paid by insurers and are widely used in research and are often used to define exposures and outcomes. However, there are concerns that pharmacy claims data may not accurately capture several types of treatments—such as over-the-counter medications, free drug samples, or low-cost generic programs that don't go through insurance claims. Brookhart illustrated this issue with new users of statins—where there were outcomes that were influenced by patients taking samples of medication prior to having their prescriptions filled. The effect of treatment was masked by the sample drug impact, the appropriate baseline was mis-specified, and potentially early adverse events were missed. Brookhart noted the potential for further examination of patterns of care prior to initiation of treatment and a combination study of EHR with claims data. He also described an example of warfarin prescriptions that may be missing from claims databases as alternatives became more available.

The discussion that followed yielded the following points and comments:

- Triangulation among multiple data sources and using multiple approaches is important. Most of these issues can be addressed through thoughtful modification of study design, and linkage of additional data. The methods of triangulation need to be easily accessible, applicable, usable, and described.

- Health researchers could think about utilizing non-traditional data sources to fill in missing information.
- Resources are really important to consider; higher quality data quality checks can be expensive and slow. Rapid turnaround and quality seem to be at a tradeoff.
- A consequence is that do we need to re-evaluate outcome definitions. Validation studies for outcomes would be useful.
- There is a lot of value in linking data sets; however, there are policy challenges and methods to do this kind of work that will need to be addressed. Going to the patient is one way to consolidate different data sources (as is done in PPRNs). However, there are often still large amounts of data missing after data are linked.
- There are some study design mechanisms that have been developed to help address missing data (Sturmer 2013).
- Recommendations to evaluate the quality of data have to be aligned with the frequency of queries being made of network systems.
- Probability sampling could be applied, using quality checks on a subset of a sample and then projecting the findings to the rest of the sample using the sampling weights. Sampling could be used to estimate measurement error and then analytic methods could be used to account for measurement error.
- Assessing temporal trends in outcomes, confounders, and exclusion criteria is essential, and we would expect that any paper should include the background rate of the outcome of interest, independent of exposure, over time. This information should be presented to ensure that changes that naturally occur over time—due to changes in policy and data generation—are not confused with the effect of an intervention. Validation of outcomes is also pivotal as definitions change over time.

Related resources:

- Lauffenburger, Julie C., et al. "Completeness of prescription information in US commercial claims databases." Pharmacoepidemiology and drug safety22.8 (2013): 899-906. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4012425/
- Schafer, Joseph L. (2008). NORM: Analysis of Incomplete Multivariate Data under a Normal Model, Version 3. Software Package for R. University Park, PA: The Methodology Center, the Pennsylvania State University.

Inconsistencies in Common Data Model Field Use

Jeff Brown from Harvard Pilgrim Health Care Institute/ Harvard Medical School next presented on inconsistencies in common data model field use. The presentation illustrated the example of HL7 (a format to transmit data) null flavors—where there are 16 different ways to say that the value is unknown (e.g., not applicable, unknown, not asked, not present, etc.). Brown noted that the importance of missingness is context specific—and data must be reviewed by each researcher to determine what constitutes "meaningful" missingness for the research question being studied. The study design and purpose should dictate whether the missingness matters—for example, it wouldn't matter that many flu vaccinations are not observed if that misclassification would not matter for a specific study. He noted that understanding data capture context is key for interpreting the data generated from a specific health system, mentioning that although EMR data may seem to include information on topics like indications for prescriptions, the information is often not reliable due to specific local policies and workflows. It is a risk to make data easily available because the data might be

misunderstood and therefore misused if the context of the data collection processes isn't known. A component of PCORnet is that it will examine common data model variables, formats, time trends, categorical variables (examined using frequencies), continuous variables (examined using distributions), and provide outputs at the summary level for individual data partners and across multiple data partners. Marked differences across data partners may indicate different data capture processes.

The discussion that followed yielded the following points and comments:

- There are other tools, like heat maps or other graphics, which could be used to enhance broad data visualization and help researchers understand the background of the data. Researchers might consider using visualizations to help understand the data for both the target population and more general populations.
- There is an opportunity to develop terminology for characteristics of data capture that can be described generally and shared to help inform analysis. There could be better communication of the bounds of what datasets can be useful for.
- It is important to understand how clinicians interact with EMRs; they often record clinical notes into EMRs, which are then converted into codes by the EMR. Shifts in EMR systems can lead to changes in how a condition is defined. There was discussion about whether these changes should be annotated or corrected. It should be noted that there was no discussion as to the complexity that will arise in incorporating ICD-10 codes into analysis, although it was noted that some investigators have already started to use data with both ICD-9 and ICD-10 coding.
- It is difficult to determine how scalable visualizations are. There are opportunities with software to examine trends and inflections without manual visualizations (such as software that can detect inflection points without using visualizations).
- PCORnet will offer opportunities to study differences in data capture and standardization and what they mean for the broader research field.
- There are also opportunities to share meta-data and lessons learned from the use of variables or in the cases of changes in processes/definitions. Audit logs are available in EMRs, and there has not been much analysis of these data to date. There are opportunities to explore pilots in the analysis of audit logs and chart review functions in EMRs. It is also interesting to examine when clinicians record data—the further from the time of the event, the less reliable the records become. The number of people that use a chart is also often an indicator of reliability.

Related resources:

- Meredith N. Zozus; W. Ed Hammond; Beverly B. Green; Michael G. Kahn; Rachel L. Richesson,; Shelley A. Rusincovitch; Gregory E. Simon; Michelle M. Smerek. Assessing Data Quality for Healthcare Systems Data Used in Clinical Research (Version 1.0)

Informative Presence of Data (Related to Health Status) – An Extended Form of Selection Bias

Bob Glynn, from Brigham & Women's Hospital, presented on the informative presence of data due to variable subject surveillance. RCTs are currently seen as the paradigm for causal efficacy and CER, but there are limitations in determining what comparators should be. For example, placebo controls and non-user referents are potential options, but both have limitations. Glynn suggested that new user designs could address some of these limitations, especially because the confounders affecting initiation may differ from those affecting persistence. Glynn discussed the characteristics of variable surveillance in effectiveness/safety studies for exposure, outcome, and covariate assessment. In summary, Glynn

discussed the key issues, including that differential surveillance of alternatively treated subjects is a clear challenge to valid causal inference in observational research. He cautioned that if differential surveillance is not addressed in the study design stage, valid inference may be impossible. New user designs, with active comparators and consideration of the time scale of treatments can help. He also suggested that restriction to subjects with comparable treatment propensities is another promising strategy.

The discussion that followed yielded the following points and comments:

- Researchers should have a focus on design over addressing missing data through analytic methods.
- But when analysis is necessary, there is potential for standards. Which techniques should be used when?
- Additional research, including theoretical and simulation evaluation, is needed to provide some indication of how strong associations with surveillance patterns need to be to cause bias and under what circumstances (i.e., which estimands).

Related resources:

- Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement Error in Nonlinear Models: A Modern Perspective, 2nd ed. Chapman and Hall CRC Press; 2006.
- Little RJ et al. The prevention and treatment of missing data in clinical trials. N Engl J Med 2012; 367:1355-1360.

Common Practices for Patient Level Linkage

Adam Wilcox, member of the PCORI Methodology Committee, then introduced the session on patient level linkage. Wilcox discussed the resources needed to maintain an enterprise master patient index (EMPI), and the error rates of such systems, highlighting that most errors are false negatives (not linking two records from the same patient), and that linkage is more problematic for different populations, and often is associated with populations that already have health disparities for other reasons. There are limitations in patient data integration, especially relating to possible re-identification, which can be done with high accuracy with only a few data fields.

Related resources:

- A simple heuristic for blindfolded record linkage. Journal of the American Medical Informatics Association. Weber, S. C., Lowe, H., Das, A., Ferris, T. 2012; 19 (E1): E157-E161

Inconsistent Data and Missingness in Linking Types of Data within and between EMR and Claims

Keith Marsolo, from Cincinnati Children's Hospital, presented on inconsistent data and missingness in linking types of data within and between EMR and claims. He noted that there is limited research in this area, and few institutions are able to effectively link data given legal and regulatory hurdles. Marsolo discussed differences in what is defined as a clinical encounter. In claims systems, encounters are defined as interaction with providers where an insured person receives service; however, in EMRs, these are defined as significant events in the life of the record (i.e., each contact with the patient). Within EMRs, this varies depending on the kind of practice, vendor, and installed functionality. He stated that there are several reasons for needing to link claims with EMR data, such as discovering/confirming the existence of an event or exposure.

The discussion that followed yielded the following points and comments:

- There was discussion about the need for analytical constructs that roll up or integrate events from multiple sources. Marsolo suggested that these could be similar to Observational Medical Outcomes Partnership (OMOP) era tables or related to episodes of care. This led to questions about how to attribute specific encounters and the need to distinguish between study-specific data linkage and linkage as a larger part of a data load.

Missing Longitudinal Data (Outcomes, Exposures, Time Varying Covariates)

Lesley Curtis, from Duke Clinical Research Institute, discussed missing longitudinal data in the EHR/EMR and the magnitude of that problem. Curtis described the importance of understanding longitudinal data. For example, complete capture is less important for health states that are stable (e.g., weight) or care events that leave an enduring signal (e.g., CABG). However, complete capture is very important for health states that are labile (e.g., hemoglobin) or care events that leave only a transient signal (e.g., suicide attempt). She stated that missing data are less common when the exposure and outcome occur in the same care setting and close in time (such as comparing in-hospital mortality for alternative bariatric procedures) and more common when the exposure and outcome occur in different care settings and/or different time periods (such as comparing the incidence of anemia over two years for alternative bariatric procedures). Curtis described insights from linking payer claims with lab groups, using an example where CMS data were linked with tests from LabCorp, spanning 2.4 million Medicare beneficiaries across 10 states.

The discussion that followed yielded the following points and comments:

- There are clear trade-offs between restricting samples and acquiring more data; this was illustrated by the example of linking labs with claims data, where there is variation in the laboratory provider market by state. Acquiring more data can often require a great deal more resources.
- It is not straightforward to implement current statistical approaches to handling missing longitudinal data in this setting where observations do not occur on a scheduled time basis, but when a patient seeks care. Making multiple imputation methods and software applicable to settings with irregular (and potentially informative) observation times is an opportunity for further research.

Related resources:

- Bynum JPW, Bernal-Delgado E, Gottlieb D, Fisher E. Assigning Ambulatory Patients and Their Physicians to Hospitals: A Method for Obtaining Population-Based Provider Performance Measurements. Health Services Research. 2007;42(1 Pt 1):45-62. doi:10.1111/j.1475-6773.2006.00633.x.
- Hammill BG, et al. Linkage of Laboratory Results to Medicare Fee-for-Service Claims. Med Care. 2015 Nov;53(11):974-9

Research Questions and Themes

After presentations from workgroup members, the workgroup was asked to discuss areas where researchers using EHR or claims data sets might benefit from guidance or standards. These topics included:

- Workgroup members identified the need for:
    - A common nomenclature for data quality
    - Metrics for what might be considered dirty data are needed – what is "bad"? What is "good"? These definitions might be use-specific (Fitness for Use).
    - Minimal set of steps for "cleaning" before the use of EHR and/or claims data and a common way to describe these steps for data users.
    - Definition of an adequate data quality report
    - Metadata learnings and training for PCORnet
- For research questions, there was consensus among the group that researchers will need to:
    - Understand the mechanism of the missingness
    - Understand how the data are generated and any underlying incentives. Develop a terminology for describing data collection contexts.
    - Evaluate whether addressing missing data would significantly change the findings of research (Fitness for Use).
- In analysis, researchers will need to:
    - Use different approaches to triangulate estimates
    - Always look at the background rate of outcome of interest over time, without regard to treatment
    - Multiple imputation probably be appropriate in most cases if have quality data
    - Perform sensitivity analyses over a range of assumptions
- In dissemination, publications should describe the project's data cleaning approach in detail in the Methods section.

The workgroup then discussed the research questions that, if answered, would benefit network-based research, like PCORnet.

Research topics that emerged from the discussion were:

General ideas:

- PCORnet could be used as a testbed for pilot projects that collaborate across different CDRNs/PPRNs, potentially with academic methodological investigators.
- Probability sampling could be used to check the quality of a subset of the sample, and the quality of the remainder could be predicted using the sampling weights.
    - Additionally, measurement error techniques could be used to incorporate estimates of data quality into a data analysis.
- Sensitivity analyses should be conducted under a range of assumptions to understand the impact of assumptions and their violation.
- Validation of outcomes with linked EMRs and claims is needed—this may be needed both globally (across the network) and at the individual site level, depending on the outcome of interest
- Tools need to be provided for PCORnet sites for cleaning and summarizing data and for understanding the bounds of the data in order to produce standardized reports and graphics. Training also needs to be provided for this work. Some of these tools could be adapted from existing publicly available tools, but it will take time and effort to develop/adapt them, appropriately test them, and roll them out. This will require funding.

Research questions:

- Linkage issues:
  - What are the privacy and policy issues for linkage and how can they be addressed?
  - What is the appropriate trade-off between false positive rates and false negative rates in linkage? These have different implications for IRBs and privacy and marked differences in analysis results.
  - How can patient level linkage methods be improved in the face of missing or poor quality data?[2]
- Understanding missingness
  - How does missingness vary across sites and delivery models?
  - Can pattern mixture models be applied to model dropout versus no dropout?
  - What missingness occurs in covariates? Does this missingness meet the assumption of MCAR (missing completely at random)?
  - Should a pilot be done of the utility of EMR audit logs in order to understand the workflow-related causes of missing data?
  - Can simulation models be used to identify how much missingness causes problems (preferably based on the covariance structure of PCORnet key covariates in existing cohorts)
  - Can simulated populations be used to understand missingness and its impact?
  - Additional research, including theoretical and simulation evaluation, is needed to provide guidance on how strong associations with surveillance patterns need to be to cause bias and under what circumstances (i.e., which estimands).
- Imputation
  - Should edit imputation and analytic edit approaches be combined and, if so, how?
  - How can multiple imputation be made user-friendly? When should it be applied and how in the EMR setting? How should it be applied to 3MM records? In various clinical scenarios?
  - Can Gelman's multiple imputation software be applied on a large scale in PCORnet?
  - Can software be developed for multiple imputation of unstructured longitudinal data?
  - Making multiple imputation methods and software applicable to settings with irregular (and potentially informative) observation times is an opportunity for further research.
- General analytic considerations:
  - Should PCORI pilot the use of PCORnet by several researchers for specific research questions before trying to make it broadly available?
  - Should efforts be undertaken to validate outcomes that have been validated by Sentinel in claims, but revalidate them in EHR linked with claims as needed, to provide a minimal set of outcomes that have been "pre-validated" in PCORnet?
- Open questions:
  - How will ICD-9 conversion to ICD-10 be handled in PCORnet? Are pilots needed, particularly for longitudinal studies?

[2] Ong TC, Mannino MV, Schilling LM, Kahn MG. Improving record linkage performance in the presence of missing linkage data. J Biomed Inform. 2014 Dec;52:43–54.

Next Steps

The workgroup's recommendations focus on achievable next steps to improve methods to minimize and handle missing data in EHR/EMR and claims data. The group, led by the Network Research Methods Sub-Committee of the PCORI Methodology Committee has discussed several next steps for this effort.

Next steps for this work stream are planned to include:

- **Mapping of general guidance topics to methodology standards**
    - Examine ways to provide guidance topics on handling missing data to supplement and enforce the PCORI Methodology Standards
    - Consider areas where new Methodology Standards should be developed.
- **Convening mini-workshops/webinars to further develop topics**
    - Topics to include methodologies for assessing missing data and analysis methods for datasets containing missing data.
    - Bringing together researchers that have successfully utilized EMR data in the past with researchers that are new to this data source or preparing for PCORnet studies would be a good use of PCORI resources if PCORI would like PCORnet to be broadly utilized.
- **The Development of Formal Recommendations**
    - Recommendations on conducting pilots for data quality within PCORnet
    - Potential PFA development around identified research questions